

Application of Information Theory, Lecture 8

Kolmogorov Complexity and Other Entropy Measures

Handout Mode

Iftach Haitner

Tel Aviv University.

May 17, 2018

Part I

Other Entropy Measures

Other entropy measures

Let $X \sim p$ be a random variable over \mathcal{X} .

- ▶ Recall that **Shannon entropy** of X is
$$H(X) = \sum_{x \in \mathcal{X}} -p(x) \cdot \log p(x) = \mathbb{E}_X [-\log p(X)]$$
- ▶ **Max entropy** of X is $H_0(X) = \log |\text{Supp}(X)|$
- ▶ **Min entropy** of X is $H_\infty(X) = \min_{x \in \mathcal{X}} \{-\log p(x)\} = -\log \max_{x \in \mathcal{X}} \{p(x)\}$
- ▶ **Collision probability** of X is $\text{CP}(X) = \sum_{x \in \mathcal{X}} p(x)^2$
Probability of collision when drawing two independent samples from X
- ▶ **Collision entropy/Renyi entropy** of X is $H_2(X) = -\log \text{CP}(X)$
- ▶ For $\alpha \neq 1 \in \mathbb{N}$ — $H_\alpha = \frac{1}{1-\alpha} \log \left(\sum_{i=1}^n p_i^\alpha \right) = \frac{\alpha}{1-\alpha} \log(\|p\|_\alpha)$
- ▶ $H_\infty(X) \leq H_2(X) \leq H(X) \leq H_0(X)$ (Jensen)
Equality iff X is uniform over \mathcal{X}
- ▶ For instance, $\text{CP}(X) \leq \sum_x p(x) \max_{x'} p(x') = \max_{x'} p(x')$. Hence,
 $H_2(X) \geq -\log \max_{x'} p(x') = H_\infty(X)$.
- ▶ Claim: $H_2(X) \leq 2 \cdot H_\infty(X)$
- ▶ Proof: $\text{CP}(X) \geq (\max_{x'} p(x'))^2$. Hence, $-\log \text{CP}(X) \leq -2 H_\infty(X)$

Other entropy measures, cont

- ▶ No simple chain rule.
- ▶ Let $X = \perp$ wp $\frac{1}{2}$ and uniform over $\{0, 1\}^n$ otherwise, and let Y be indicator for $X = \perp$.
- ▶ $H_\infty(X|Y = 1) = 0$ and $H_\infty(X|Y = 0) = n$. But $H_\infty(X) = 1$.

Sample entropy

For random variable X , define $H_X(x) = -\log \Pr[X = x]$.

- ▶ $H(X) = \mathbb{E}_{x \leftarrow X} H_X(x)$
- ▶ $H_\infty(X) = \min_{x \in X} H_X(x)$

Section 1

Shannon to min entropy

Shannon to Min entropy

Given rv $X \sim p$, let X^n denote n independent copies of X , and let $p^n(x_1, \dots, x_n) = \prod_{i=1}^n p(x_i)$.

Lemma 1

Let $X \sim p$ and let $\varepsilon > 0$. Then $\Pr[-\log p^n(X^n) \leq n \cdot (H(X) - \varepsilon)] < 2 \cdot e^{-2\varepsilon^2 n}$.

- Compare to $-\log p^n(\mathbf{x}) \leq n \cdot H_\infty(X)$, for any $\mathbf{x} \in \text{Supp}(X)$

Corollary 2

\exists rv W that is $(2 \cdot e^{-2\varepsilon^2 n})$ -close to X^n , and $H_\infty(W) \geq n(H(X) - \varepsilon)$.

Proof: $W = X^n$ if $X^n \in A_{n,\varepsilon}$, and “well spread” outside $\text{Supp}(X^n)$ otherwise.

Shannon to min entropy, proof

$$p^n(x_1 \dots, x_n) = \prod_{i=1}^n p(x_i).$$

Lemma 3

Let $X \sim p$ and let $\varepsilon > 0$. Then $\Pr[-\log p^n(X^n) \leq n \cdot (H(X) - \varepsilon)] < 2 \cdot e^{-2\varepsilon^2 n}$.

Proof: (quantitative) AEP.

Proposition 4 (Hoeffding's inequality)

Let Z^1, \dots, Z^n be iids over $[0, 1]$ with expectation μ . Then,

$$\Pr\left[\left|\frac{\sum_{j=1}^n Z^j}{n} - \mu\right| \geq \varepsilon\right] \leq 2 \cdot e^{-2\varepsilon^2 n} \text{ for every } \varepsilon > 0.$$

- ▶ $A_{n,\varepsilon} := \{\mathbf{x} \in \text{Supp}(X^n) : 2^{-n(H(X)+\varepsilon)} \leq p^n(\mathbf{x}) \leq 2^{-n(H(X)-\varepsilon)}\}$
- ▶ $-\log p^n(\mathbf{x}) \geq n \cdot (H(X) - \varepsilon)$ for any $\mathbf{x} \in A_{n,\varepsilon}$
- ▶ Taking $Z_i = -\log p(X_i)$ and $\mu = H(X)$, it follows that

$$\Pr[X^n \notin A_{n,\varepsilon}] = \Pr\left[\left|n\mu - \sum_i Z_i\right| \geq \varepsilon n\right] = \Pr\left[\left|\mu - \frac{\sum_i Z_i}{n}\right| \geq \varepsilon\right] \leq 2 \cdot e^{-2\varepsilon^2 n}. \square$$

Shannon to Min entropy, conditional version

Lemma 5

Let $(X, Y) \sim p$ let $\varepsilon > 0$. Then

$$\Pr_{(x^n, y^n) \leftarrow (X, Y)^n} \left[-\log p_{X^n|Y^n}^n(x^n|y^n) \leq n \cdot (H(X|Y) - \varepsilon) \right] < 2 \cdot e^{-2\varepsilon^2 n}.$$

Proof: same proof, letting $Z_i = \log p_{X|Y}(X_i|Y_i)$

Corollary 6

\exists rv W over $\mathcal{X}^n \times \mathcal{Y}^n$ such that

- ▶ $\text{SD}(W, (X, Y)^n) \leq 2 \cdot e^{-2\varepsilon^2 n}$.
- ▶ $\text{SD}(W_{\mathcal{Y}^n}, Y^n) = 0$, and
- ▶ $H(W_{\mathcal{X}^n} | W_{\mathcal{Y}^n} = \mathbf{y}) \geq n \cdot (H(X|Y) - \varepsilon)$, for any $\mathbf{y} \in \text{Supp}(Y^n)$

Proof: ?

Section 2

Renyi-entropy to Uniform Distribution

Extraction

Goal: given a random variable over $\{0, 1\}^n$, with k bits of “entropy”, extract close to k uniform bits.

- ▶ Let $U_t \sim \{0, 1\}^t$.
- ▶ Deterministic extractors:
 $\text{Ext}: \{0, 1\}^n \mapsto \{0, 1\}^m$, such that for any X over $\{0, 1\}^n$ with $H_\infty(X) \geq K$ it holds that $\text{SD}(\text{Ext}(X), U_m) \leq \varepsilon$.
- ▶ Impossible to achieve even for $k = n - 1$ and $m = 1$!
- ▶ Seeded extractors:
 $\text{Ext}: \{0, 1\}^n \times \{0, 1\}^d \mapsto \{0, 1\}^m$, such that for any X over $\{0, 1\}^n$ with $H_\infty(X) \geq K$ it holds that $\text{SD}((\text{Ext}(X, U_d), U_d), (U_m, U_d)) \leq \varepsilon$.
- ▶ Typically, we would like d to be as small as possible.
- ▶ Very useful concept

Pairwise independent hashing

Definition 7 (pairwise independent function family)

A function family $\mathcal{G} = \{g: \mathcal{D} \mapsto \mathcal{R}\}$ is **pairwise independent**, if $\forall x \neq x' \in \mathcal{D}$ and $y, y' \in \mathcal{R}$, it holds that $\Pr_{g \leftarrow \mathcal{G}} [g(x) = y \wedge g(x') = y'] = (\frac{1}{|\mathcal{R}|})^2$.

- ▶ Example: for $\mathcal{D} = \{0, 1\}^n$ and $\mathcal{R} = \{0, 1\}^m$ let $\mathcal{G} = \{(A, b) \in \{0, 1\}^{m \times n} \times \{0, 1\}^m\}$ with $(A, b)(x) = A \times x + b$. (additions are over $\mathbb{F}(2)$, e.g., \oplus)
- ▶ 2-universal families: $\Pr_{g \leftarrow \mathcal{G}} [g(x) = g(x')] = \frac{1}{|\mathcal{R}|}$.
- ▶ Example for universal family that is not pairwise independent?
- ▶ Many-wise independent

Leftover hash lemma

Lemma 8 (leftover hash lemma)

Let X be a rv over $\{0, 1\}^n$ with $H_2(X) \geq k$, let $\mathcal{G} = \{g: \{0, 1\}^n \mapsto \{0, 1\}^m\}$ be 2-universal, and $G \leftarrow \mathcal{G}$. Then $SD((G, G(X)), (G, \sim \{0, 1\}^m)) \leq \frac{1}{2} \cdot 2^{(m-k)/2}$.

Lemma 9

Let p be dist. over \mathcal{U} with $CP(p) \leq \frac{1+\delta}{|\mathcal{U}|}$, then $SD(p, \sim \mathcal{U}) \leq \frac{\sqrt{\delta}}{2}$.

Proof: Let q be the uniform distribution over \mathcal{U} .

- ▶ $\|p - q\|_2^2 = \sum_{u \in \mathcal{U}} (p(u) - q(u))^2 = \|p\|_2^2 + \|q\|_2^2 - 2\langle p, q \rangle = CP(p) - \frac{1}{|\mathcal{U}|} \leq \frac{\delta}{|\mathcal{U}|}$
- ▶ Chebyshev Sum Inequality: $(\sum_{i=1}^n a_i)^2 \leq n \cdot \sum_{i=1}^n a_i^2$
- ▶ Hence, $\|p - q\|_1^2 \leq |\mathcal{U}| \cdot \|p - q\|_2^2$
- ▶ Thus, $SD(p, q) = \frac{1}{2} \|p - q\|_1 \leq \frac{\sqrt{\delta}}{2} \cdot \square$

To deuce the proof of **Lemma 8**, we notice that

$$CP(G, G(X)) \leq \frac{1}{|\mathcal{G}|} \cdot (2^{-k} + 2^{-m}) = \frac{1+2^{m-k}}{|\mathcal{G}| \cdot 2^m} = \frac{1+2^{m-k}}{|\mathcal{G} \times \{0, 1\}^m|}$$

Part II

Kolmogorov Complexity

Description length

- ▶ What is the description length of the following strings?
 1. 01
 2. 011010100000100111100110011001111110
 3. 11101010011000110011110001010101111
- ▶
 1. Eighteen 01
 2. First 36 bit of the binary expansion of $\sqrt{2} - 1$
 3. Looks random, but 22 ones out of 36
- ▶ Bergg's paradox: Let s be "the smallest positive integer that **cannot** be described in twelve English words"
- ▶ The above is a definition of s , of less than twelve English words...
- ▶ Solution: the word "described" above in the definition of s is not well defined

Kolmogorov complexity

- ▶ For a string $x \in \{0, 1\}^*$, let $K(x)$ be the length of the **shortest** C^{++} program (written in binary) that outputs x (on empty input)
- ▶ Now the term “described” is well defined.
- ▶ Why C^{++} ?
- ▶ All (complete) programming language/computational model are essentially equivalent.
- ▶ Let $K'(x)$ be the description length of x in another complete language, then $|K(x) - K'(x)| \leq \text{const}$.
- ▶ What is $K(x)$ for $x = \underbrace{0101010101 \dots 01}_{n \text{ pairs}}$
- ▶ “For $i = 1 : i^{++} : n$; print 01”
- ▶ $K(x) \leq \log |X| + \text{const}$
- ▶ This is considered to be small complexity. We typically ignore $\log n$ factors.
- ▶ What is $K(x)$ for x being the first n digits of π ?
- ▶ $K(x) = \log |x| + \text{const}$

More examples

- ▶ What is $K(x)$ for $x \in \{0, 1\}^n$ with j ones?
- ▶ Recall that $\binom{n}{j} \leq 2^{n \cdot h(j/n)}$
- ▶ Hence $K(x) \leq \log n + n \cdot h(j/n)$

Bounds

- ▶ $K(x) \leq |x| + \text{const}$
- ▶ Proof: “output x ”
- ▶ Most sequences have high Kolmogorov complexity:
- ▶ At most 2^{n-1} (C^{++}) programs of length $\leq n - 1$
- ▶ 2^n strings of length n
- ▶ Hence, at least $\frac{1}{2}$ of n -bit strings have Kolmogorov complexity at least n
- ▶ In particular, a random sequence has Kolmogorov complexity $\approx n$

Applications

(another) Proof that there are infinity many primes.

- ▶ Assume there are finitely many primes p_1, \dots, p_m
- ▶ Any length n integer x can be written as $x = \prod_{i=1}^m p_i^{d_i}$
- ▶ $d_i \leq n$. Hence $|d_i|$, length of d_i , $\leq \log n$
- ▶ Hence, $K(x) \leq m \cdot \log |X| + \text{const}$
- ▶ But for most n -bit numbers, $K(x) \geq n$

Conditional Kolmogorov complexity

- ▶ $K(x|y)$ — Kolmogorov complexity of x given y . The length of the shortest program that outputs x on input y

- ▶ Chain rule

$$K(x, y) \approx K(y) + K(x|y)$$

H vs. K

$H(X)$ speaks about a random variable X and $K(x)$ of a string x , but

- ▶ Both quantities measure the amount of uncertainty or randomness in an object
- ▶ Both measure the number of bits it takes to describe an object
- ▶ Another property: Let X_1, \dots, X_n be iid, then whp
$$K(X_1, \dots, X_n) \approx H(X_1, \dots, X_n) = nH(X_1)$$
- ▶ Proof: ? AEP
- ▶ Example: coin flip $(0.7, 0.3)$ then whp we get a string with
$$K(x) \approx n \cdot h(0.3)$$

Computability of K

- ▶ Can we compute $K(x)$?
- ▶ Answer, No.
- ▶ Proof: Assume K is computable by a program of length C
- ▶ Let s be the smallest positive integer s.t. $K(s) > C + 1000$
- ▶ s can be computed by the following program:
 1. $x = 0$
 2. While $(K(x) < C + 1000)$: $x++$
 3. Output x
- ▶ Thus $K(s) < C + \log 1000 + \text{const} < C + 1000$
- ▶ Bergg's Paradox, revisited:
- ▶ s — the smallest positive number with $K(s) > 1000$
- ▶ This is not a paradox, since the description of s is not short.

Explicit large complexity strings

- ▶ Can we give an explicit example of string x with large $k(x)$?

Theorem 10

\exists constant C s.t. the theorem $K(x) \geq C$ *cannot* be proven (under any reasonable axiom system).

- ▶ For most strings $K(x) > C + 1$, but it cannot be proven even for a single string
- ▶ $K(x) \geq C$ is an example for a theorem that cannot be proven, and for most x 's cannot be disproved.
- ▶ Proof: for integer C define the program T_C :
 1. $y = 0$
 2. If $y = (x, z)$ and z is a proof for the statement $k(x) > C$. Output x .
 3. y^{++}
- ▶ $|T_C| = \log C + D$, where D is a const
- ▶ Take C such that $C > \log C + D$
- ▶ If T_C stops and outputs x , then $k(x) < \log C + D < C$, a contradiction to the fact that \exists proof that $k(x) > C$.

Universal compression

- ▶ A program of length $K(x)$ that outputs x , compresses x into $k(x)$ bit of information.
- ▶ Example: length of the human genome: $6 \cdot 10^9$ bits
- ▶ But the code is redundant
- ▶ The relevant number to measure the number of possible values is the Kolmogorov complexity of the code.
- ▶ No-one knows its value...